**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 6.199**

***IJCSMC, Vol. 8, Issue. 4, April 2019, pg.136 – 141***

# The Use of Data Mining Techniques in Heart Disease Prediction

**Dr. Yasemin Gultepe**
Computer Engineering Department, Kastamonu University, Turkey
yasemingultepe@kastamonu.edu.tr

**Sabah Rashed**
Material Sciences and Engineering, Kastamonu University, Turkey
sabahrashed1984@gmail.com

***Abstract— One-third of deaths worldwide are the result of heart disease, the rate of death from heart disease is higher than the mortality rates due to cancer. The cause of these deaths is the lack of knowledge of the symptoms of this disease or lack of attention to these symptoms. Where the patient believes that these symptoms due to fatigue or other diseases less serious. And as a result of the enormous amounts of data in the field of heart disease and the corresponding development in the field of computing and the availability of data processing programs. It becomes easy now to use these programs to predict heart disease. In this article we used Weka software as one of Data Mining techniques in heart disease prediction by testing heart-c.arff dataset obtained from UCI repository against several data classification techniques using naïve bayes and J84 classification algorithm.***

***Keywords— Heart Disease, Heart Disease prediction, Data Mining, naïve bayes classification algorithm, J84 classification algorithm.***

## I. INTRODUCTION

Heart disease is one of the most widespread diseases these days. The prevalence of these diseases is due to many reasons, including age, nature of work, genetic causes, alcohol, drugs, smoking, high blood pressure, high cholesterol and obesity. Symptoms of heart disease such as fatigue and chest pain are similar to those caused by stress due to excessive work or lack of sleep. So most people with these symptoms think that they are caused by fatigue and do not check the heart to ensure their safety from heart disease until they faced a real heart problem. Reviewing historical data on heart disease is almost impossible due to the vast amount of these data. But the use of computing made it easy to conduct searches in these data and extract useful information and important statistics from these data. One of the important computing techniques is Data Mining which is used frequently to deal with the medical historical data. Data Mining contains of many classification algorithms like naïve bayes and J84. These algorithms are used to predict heart diseases.

The rest of the paper has been organized as follows. Section 2 reviews the previous related works of heart disease prediction. Section 3 defines the simulation technique called Weka 3.7.9. Section 4 contains the conclusion of our work.

 

## II. RELATED WORKS

An attempt to predict heart disease by analysing historical medical data has been discussed by many authors. Each one of them try to use different methodology in his article. Reference [1] tries to predict heart disease by using medical data (age, sex, blood pressure and blood sugar), by mining these data using two methods (Neural Network, K-Means Clustering). And he found that Artificial Neural Networks outperform K Means clustering in all the parameters. Another comparison founded in [2], where the author used (13) attribute structured clinical database from UCI Machine Learning Repository in prediction by using another two methods which are (Decision tree and Naive Bayes). The author found that the performance of Naive Bayes is better than Decision tree. The author of article [3] has increase the attribute number where he added (obesity and smoking) and also he used three prediction methods, these methods are (Neural Networks, Naive Bayes, and Decision Trees). The author found that the performance of these methods is 100%, 99.62%, and 90.74% respectively. So he agree with [2] that Neural Networks is the highest performance method for prediction. Article [4] developed an initial model for predicting intelligent heart disease using data mining techniques called it (Intelligent Heart Disease Prediction System) (IHDPS). In this article the researcher tried to find the hidden relationship between medical items like (age, sex, blood pressure and blood sugar). IHDPS can also answer "what if" questions where old decision support systems cannot answer them.

## III. METHODOLOGY

In this article, authors tested heart-c.arff dataset that obtained from UCI repository [5] against several data classification techniques using Weka software. The main aim of this article is to predict heart disease from different attributes in this dataset. The attribute num represents the (binary) class attribute: class <50 means no disease; class >50_1 indicates increased level of heart disease. The Weka software used in this assignment Version 3.8.3, obtained from [6]. Figure 1 shows the all attributes visualization.

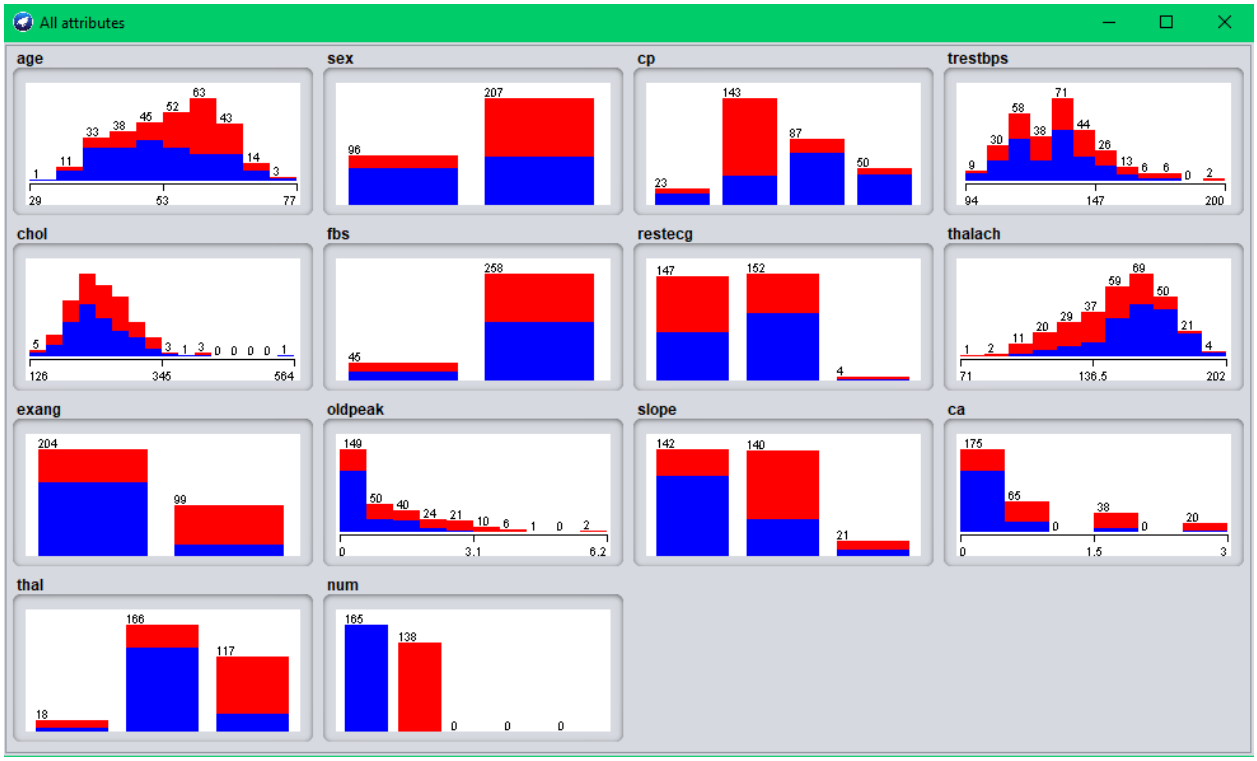


Figure.1 Attributes Visualization

### 3.1. Classifiers

From reading several published literatures, it was found that there are numerous number of classifiers used nowadays. Depending on the case of study, some classifiers outperformed others while failed in other case studies; according to [7] [8] [9]. For this assignment, we chose two of widely used classifiers, namely Naïve Bayes and Decision Tree (J48).

 

### 3.1.1. Naïve Bayes

Naïve bayes are group of probabilistic classifiers built on the Bayes' theorem [10]. Which states that:
Consider X and H
X: is an evident, X=x1, x2...,xn
H: is the hypothesis

Then

$$P(H/X) = (P(X/H)*P(H))/P(X)$$

a. **Loading Data:** Select **Explorer>Open file.** Then choose the heart disease dataset.

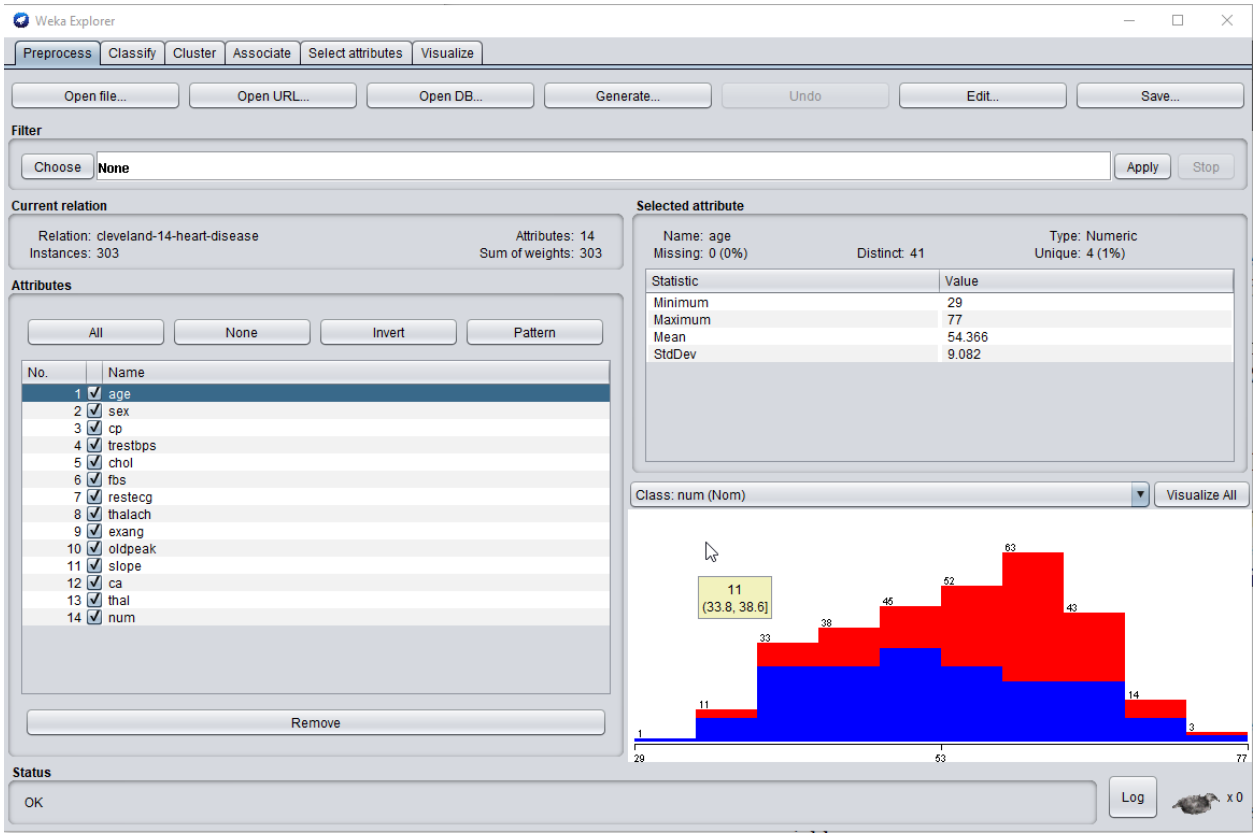


Figure.2 Loading Dataset

b. Select **Classify>Choose>Classifiers>Bayes>NaiveBayes.** Then we set the ***Percentage Split*** of data, 70% for training and 30% for testing. Figure 3 shows the obtained results from applying Naïve classifier on the selected dataset which shows 85.71% **accuracy**, 78 **correctly classified instances**, and with **Recall** value of 0.857 and **Precision** of 0.859.

 

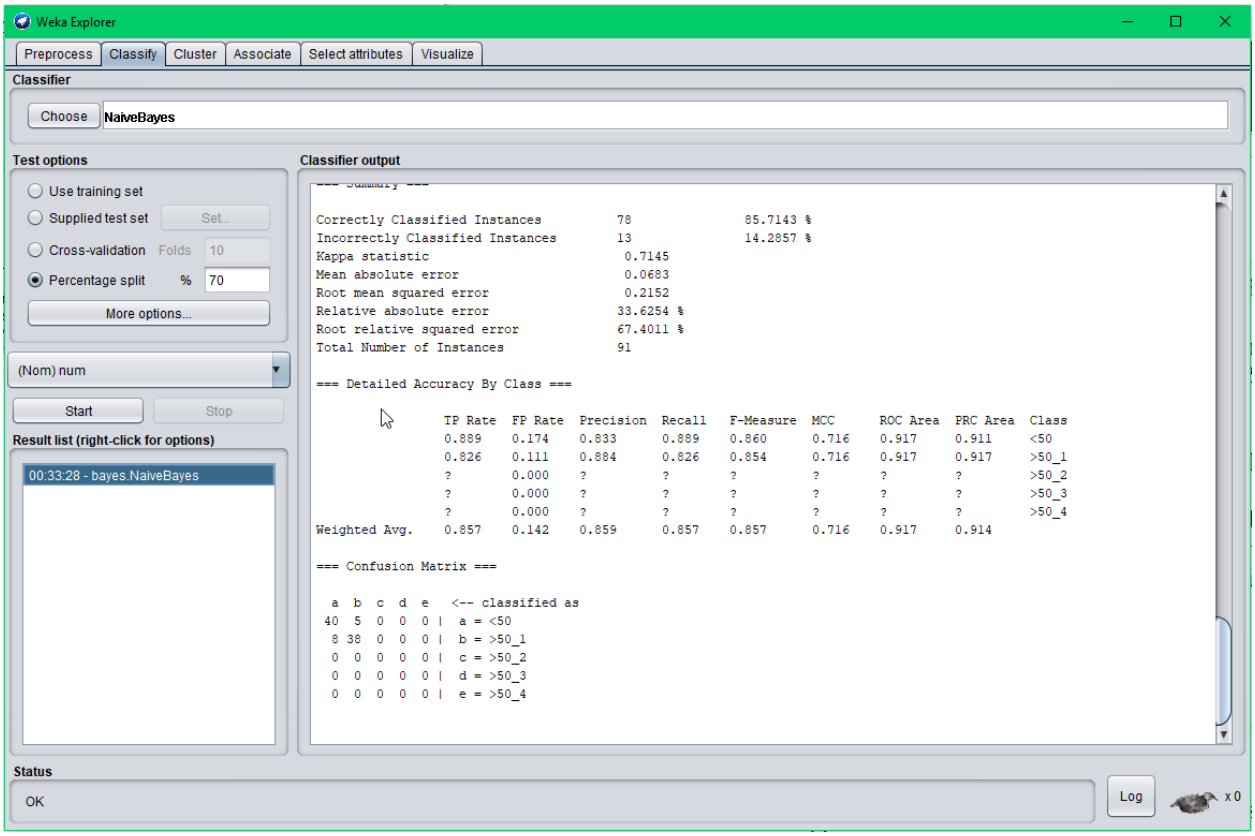


Figure.3 Naïve Bayes Classification Results

### 3.1.2 J48

J48 algorithm is a decision tree developed based on ID3 (Iterative Dichotomiser 3) by Weka as the successor of C4.5. Decision tree starts with dataset (training set), then through recursive division, it partitions the data into smaller and smaller sets to further end with sub-roots and nodes distributed over different levels. The value of these nodes and roots are identified with labels used to determine if an object/information belongs to this particular class or not [11].

J48 algorithm uses pruning method to construct decision tree leaves and branches. Pruning is a method of removing redundant data/information. Overall, it reduces the complexity and enhances the performance of classification.

### Experiment using J48

To apply J48 tree, Select **Classify>Choose>trees>J48**. Then we set the ***Percentage Split*** again - of data, 70% for training and 30% for testing. Figure 4 shows the obtained results from applying Naïve classifier on the selected dataset which shows 76.92% **accuracy**, 70 **correctly classified instances**, and with **Recall** value of 0.769 and **Precision** of 0.773.

 

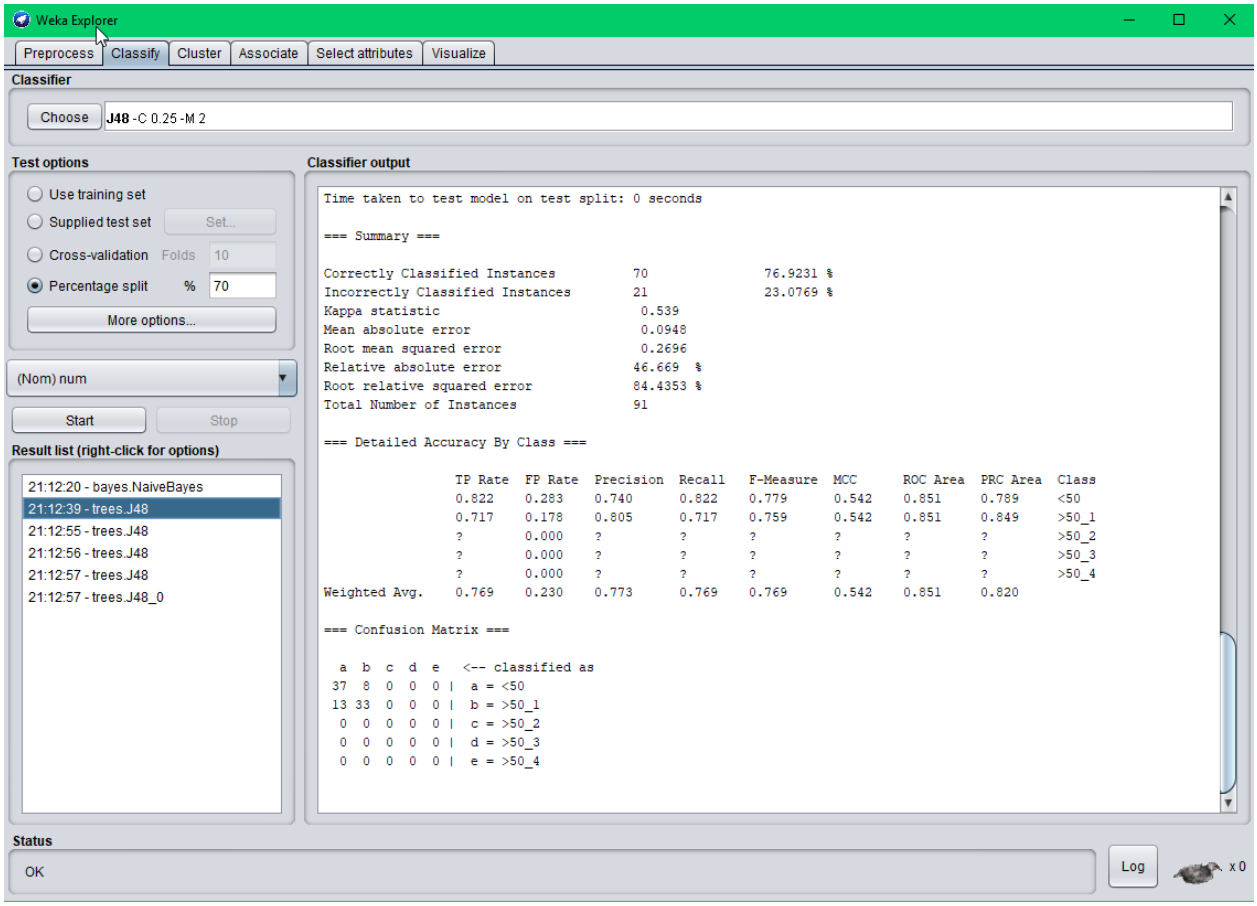


Figure.4 J48 Classification Results

### 3.2. Optimization

In Weka, Ensemble Learning is a powerful class of machine learning algorithms where instead of using the prediction of a single classifier, a set of predictions algorithms are combined and the final decision is held based on the results of the sum of prediction models. Different Ensemble models exist in Weka, such as Bagging, Boosting, and Voting.

In this article, we will try to optimize the obtained results through the use of the first ensemble algorithm, **Bagging**. From Weka main interface, Select **Classify>Choose>meta>Bagging**. Throughout several tests carried using Weka, bagging several Naïve Bayes models didn't improve the overall results. When the **numofIteration** is set to 20, 40, 60; the accuracy remained the same, 85.71. Number of Iteration defines the number of bags, i.e. the number of naïve bayes models to run concurrently with resampling.

However, optimization of J48 using Bagging increased the accuracy of classification dramatically. From Figure 4, the initial results of J48 was 76.92%. After applying Bagging with J48, we successfully reached to an accuracy of 81.31%. The table below details the tests results.

 

Table I

Bagging Results of J48

| Number of Iterations | Recall | Precision | Accuracy | No of correctly classified Instances | No of Incorrectly classified Instances |
|---|---|---|---|---|---|
| 20 | 0.78 | 0.78 | 78.02 | 71 | 20 |
| 40 | 0.802 | 0.803 | 80.21 | 73 | 18 |
| 60 | 0.813 | 0.813 | 81.31 | 74 | 17 |

## IV. CONCLUSION

In this article we tried to predict heart disease from different attributes in heart-c.arff dataset which obtained from UCI against several data classification techniques using Weka software. Where instead of using the prediction of a single classifier, a set of predictions algorithms are combined and the final decision is held based on the results of the sum of prediction models. Also we tried to optimize the obtained results through the use of the first ensemble algorithm, Bagging. Where Naïve Bayes models didn't improve the overall results. When the numofIteration is set to 20, 40, 60; the accuracy remained the same, 85.71.

However, optimization of J48 using Bagging increased the accuracy of classification dramatically. We found that the initial results of J48 was 76.92%. After applying Bagging with J48, we successfully reached to an accuracy of 81.31%. Finally we can conclude that Naïve Bayes model is less accuracy than J48.

# REFERENCES


[1] Andrea D'Souza, "Heart Disease Prediction Using Data Mining Techniques", www.ijres.org Volume 3 Issue 3 ǁ March. 2015 ǁ PP.74-77.

[2] B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining ", 2014 International Conference on Innovations in Engineering and Technology (ICIET'14).

[3] Chaitrali S. Dangare, Sulabha S. Apte, PhD., "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques ", International Journal of Computer Applications (0975 – 888), Volume 47– No.10, June 2012.

[4] Sellappan Palaniappan, Rafiah Awang,"Intelligent heart disease prediction system using data mining techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[5] Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php

[6] WEKA at https://www.cs.waikato.ac.nz/ml/weka/downloading.html

[7] Salha, et al, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction", Lecture Notes on Information Theory Vol. 2, No. 4, December 2014, PP 310-315.

[8] Vikas Chaurasia, Saurabh Pal, "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, ISSN: 2296-1739, PP 56-66.

[9] Umair, et al, "Data Mining in Healthcare for Heart Diseases", International Journal of Innovation and Applied Studies Vol. 10 No. 4 Mar. 2015, ISSN 20'28-9324, PP. 1312-1322.

[10] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[11] Kaur, Gaganjot, and Amit Chhabra. "Improved J48 classification algorithm for the prediction of diabetes." International Journal of Computer Applications 98.22 (2014).


 